Stat 310 Homework 6 Key

Chapter 5, problems 16, 18. Chapter 7, problems 3, 4, 12 (explain your answers!), 26, 33. Chapter 8, problems 2, 6, 11. Due 11/4/99.

5.16 Suppose that $X_1, \ldots, X_{20}$ are independent random variables with density functions

$$f_X(x) = 2x, \quad 0 \leq x \leq 1.$$

Let $S = X_1 + \ldots + X_{20}$. Use the Central Limit Theorem to approximate $P(S \leq 10)$.

Ok, as we add a bunch of independent and identically distributed random variables together, the distribution of the sum becomes increasingly bell-shaped (a la the CLT). Thus, to make probabilistic statements about the sum, we need to find the parameters of the approximate Normal distribution, namely μ_S and σ_S^2. That is, according to the CLT,

$$S \sim N(\mu_S, \sigma_S^2).$$

Now, as the X_i are independent,

$$\begin{aligned} E(S) &= E(\sum X_i) = \sum E(X_i) = nE(X), \\ V(S) &= V(\sum X_i) = \sum V(X_i) = nV(X). \end{aligned}$$

Here,

$$\begin{aligned} E(X) &= \int_0^1 2x^2 dx = \frac{2}{3}, \\ E(X^2) &= \int_0^1 2x^3 dx = \frac{2}{4} \\ V(X) &= E(X^2) - E(X)^2 = \frac{2}{4} - \frac{4}{9} = \frac{1}{18}. \end{aligned}$$

Thus, $S \sim N(20 * 2/3, 20 * 1/18) = N(13.33, 1.11)$. Putting it all together,

$$\begin{aligned} P(S \leq 10) &= P\left(\frac{S - 13.33}{\sqrt{1.11}} \leq \frac{10 - 13.33}{\sqrt{1.11}}\right) \\ &\approx P(Z \leq -3.16) = 0.0008. \end{aligned}$$

5.18 Suppose that a company ships packages that are variable in weight, with an average weight of 15lb and a standard deviation of 10lb. Assuming that the packages come from a large number of different customers so that it is reasonable to model their weights as independent random variables, find the probability that 100 packages will have a total weight exceeding 1700lb.

This is, not too surprisingly, another exercise in the use of the CLT. In this case, we are implicitly assuming that there is an underlying distribution of package weights, even though we don't specify the shape of that distribution. Even without the shape, we have

the center and the spread which will be enough to characterize the shape of the sum of a large number of terms from this distribution. Letting X_i denote the ith package weight, and $S = X_1 + \ldots + X_{100}$, we are trying to find $P(S \leq 1700)$. From the CLT, we know that the distribution of S is approximately normal with mean $n\mu_X$ and variance $n\sigma_X^2$, so $S \sim N(100 * 15, 100 * 10^2) = N(1500, 10000)$ (note that they gave us the standard deviation above; variances add, so we need to square this). Finally,

$$\begin{aligned} P(S \leq 1700) &= P\left(\frac{S - 1500}{\sqrt{10000}} \leq \frac{1700 - 1500}{\sqrt{10000}}\right) \\ &\approx P(Z \leq 2) = 0.9772. \end{aligned}$$

7.3. Which of the following is a random variable?

a) The population mean.

This is *not* a random variable. The mean of the distribution is a number. With respect to sampling, the value of the population mean does not change from sample to sample.

b) The population size, N.

Again, this is *not* a random variable. Numbers that define the underlying population are treated as fixed, Numbers that vary from sample to sample are treated as random variables.

c) The sample size, n.

This one is actually a bit tricky. Normally, the answer is that this is *not* a random variable; in characterizing a sampling distribution we normally speak of the distribution for samples of size n. Thus, this is a number related to the sample that does not change from sample to sample.

d) The sample mean.

Easy. This *is* a random variable. The value of $\bar{X}$ will change from sample to sample.

e) The variance of the sample mean.

The variance of the sample mean is given by σ^2/n, where σ^2 is the population variance, and hence fixed. Thus, the variance of the sample mean is actually *not* a random variable (but see part h).

f) The largest value in the sample.

This *is* a random variable; the value can change from one sample to the next.

g) The population variance.

This is *not* a random variable; σ^2 is a number.

h) The estimated variance of the sample mean.

This value, s^2, *is* a random variable, as its value can change from one sample to the next.

7.4. Two populations are surveyed with simple random samples. A survey of size n_1 is used for population I, which has a population standard deviation σ_1; a sample of size $n_2 = 2n_1$ is used for population II, which has a population standard deviation of $\sigma_2 = 2\sigma_1$. Ignoring finite population corrections, in which of the two samples would you expect the estimate of the population mean to be more accurate?

This question addresses a point raised in class, namely how much larger one has to make a sample to get a confidence interval to shrink by a given factor. In both cases here we will be using the sample mean to estimate the corresponding population mean. Consider:

$$\begin{aligned} V(\bar{X}_1) &= \sigma_1^2/n_1, \\ V(\bar{X}_2) &= \sigma_2^2/n_2 \\ &= (2\sigma_1)^2/(2n_1) = 2\sigma_1^2/n_1. \end{aligned}$$

Thus, we would expect the estimate of the population mean would be more accurate in the first sample. Remember - if you want to cut the standard deviation (proportional to the width of a confidence interval) in half, you need $2^2 = 4$ times as much data!

7.12. True or false?

a) The center of a 95% confidence interval for the population mean is a random variable.

True. The center and spread of a confidence interval will change from sample to sample. For the cases we're looking at right now, these confidence intervals for the population mean are of the form $\bar{X} \pm ks/\sqrt{n}$, where k is some constant that we choose to achieve a given level of coverage and s is the sample standard deviation. Here, the center is $\bar{X}$, which is a random variable.

b) A 95% confidence interval for μ contains the sample mean with probability .95.

False. A 95% confidence interval for μ is typically centered on the sample mean, that being our single best guess as to the value of the population mean. Hence, a 95% c.i. for μ will contain the sample mean with probability 1.

c) A 95% confidence interval contains 95% of the population.

False. The 95% refers to the fraction of times that intervals constructed from samples using this procedure will contain a specific parameter of the population. It need not contain a set fraction of the population.

d) Out of one hundred 95% confidence intervals for μ, 95 will contain μ.

False. More accurately, this is not necessarily true. Whether a given confidence interval contains μ or not depends on the sample that is drawn; this is a random quantity. In particular, this random quantity is one whose distribution we can specify! The chance that a randomly drawn sample will yield a 95% confidence interval containing μ is 0.95, so the number of intervals out of 100 that will contain μ is a binomial random variable with $n = 100, p = 0.95$.

7.26. Referring again to Example D in Section 7.3.3, suppose that a survey is done of another condominium project of 12,000 units. The sample size is 200, and the proportion planning to sell in this sample is .18.

a) What is the standard error of this estimate? Give a 90% confidence interval.

This part is an exercise in following the book's template for constructing a confidence interval for a proportion. Here, an unbiased estimator of the variance of a proportion is given by

$$s^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right).$$

The standard error is simply the square root of the sample variance, so in this case

$$s_{\hat{p}} = \sqrt{\frac{.18(1-.18)}{200-1}\left(1-\frac{200}{12,000}\right)} = 0.027.$$

A 90% confidence interval for the true proportion is given by

$$\hat{p} \pm 1.645 * s_{\hat{p}} = .18 \pm 0.0444 = (.1356, .2244).$$

b) Suppose we use the notation $\hat{p}_1 = .12$ and $\hat{p}_2 = .18$ to refer to the proportions in the two samples. Let $\hat{d} = \hat{p}_1 - \hat{p}_2$ be an estimate of the difference, d, of the two population proportions p_1 and p_2. Using the fact that $\hat{p}_1$ and $\hat{p}_2$ are independent random variables, find expressions for the variance and standard error of $\hat{d}$.

As the two quantities are independent,

$$V(\hat{d}) = V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + (-1)^2 V(\hat{p}_2) = V(\hat{p}_1) + V(\hat{p}_2).$$

Note that even though we are working with an estimate of the difference of two quantities, the variances add, not subtract. Thus,

$$\begin{aligned} s^2_{\hat{d}} &= \frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1}\left(1-\frac{n_1}{N_1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}\left(1-\frac{n_2}{N_2}\right), \\ s_{\hat{d}} &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1-1}\left(1-\frac{n_1}{N_1}\right) + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2-1}\left(1-\frac{n_2}{N_2}\right)},. \end{aligned}$$

Numerically,

$$\begin{aligned} s^2_{\hat{d}} &= \frac{.12(1-.12)}{100-1}\left(1-\frac{100}{8000}\right) + \frac{.18(1-.18)}{200-1}\left(1-\frac{200}{12000}\right) \\ &= 0.001053 + 0.000729 = 0.001783 \\ s_{\hat{d}} &= 0.04222. \end{aligned}$$

c) Since $\hat{p}_1$ and $\hat{p}_2$ are approximately normally distributed, so is $\hat{d}$. Use this fact to construct 99%, 95%, and 90% confidence intervals for d. Is there clear evidence that p_1 is really different than p_2?

Well, the form of the confidence intervals is given by

$$\hat{d} \pm (const) * s_{\hat{d}}$$

where the constants are chosen so that in the case of a normal random variable we would achieve the specified level of coverage. For the cases under consideration, these constants are 2.575, 1.96, and 1.645 for the 99%, 95% and 90% c.i.'s, respectively. To get these, consider the first one - this corresponds to $z_{0.995}$, meaning that 99.5% of the normal distribution is below this value, and by using this to define our upper bound we are excluding the 0.5% of the distribution above it. Similarly, with the lower bound we are excluding the bottom 0.5%. For the data at hand, $\hat{d} = .12 - .18 = -.06$, so the intervals are given by

$$\begin{aligned}
-.06 \pm 2.575 * (0.0422) &= (-0.1687, 0.0487), \\
-.06 \pm 1.960 * (0.0422) &= (-0.1427, 0.0227), \\
-.06 \pm 1.645 * (0.0422) &= (-0.1294, 0.0094).
\end{aligned}$$

Is there strong evidence that the population proportions are different? If they are not different, then $d = 0$. This value is included in all of the confidence intervals that we constructed for d just now, meaning that 0 is a plausible value for d and hence no, there is not strong evidence that the population proportions are different.

7.33 Suppose that in a population of N items, k are defective in some way. For example, the items might be documents a small proportion of which are fraudulent. How large should a sample be so that with a specified probability it will contain at least one of the defective items? For example, if $N = 10000$, $k = 50$, and $p = 0.95$, what should the sample size be? Such calculations are useful in planning sample sizes for accepatance sampling.

This problem is very similar to one from way back in homework 1. We attack it by figuring out the chance of not drawing any defective items in n draws in a recursive fashion:

$$P(\text{1 or more defects}) = 1 - \frac{N-k}{N} * \frac{N-k-1}{N-1} * \ldots * \frac{N-k-n+1}{N-n+1}.$$

Keep increasing n until the answer exceeds 0.95. We suggested two approximations to the final answer, based on assuming independence results:

$$\begin{aligned}
\left(\frac{N-k}{N}\right)^n &< 0.05, \\
n &\approx \log(0.05)/\log((N-k)/N) \\
\left(\frac{N-n}{N}\right)^k &< 0.05 \\
n &\approx N - \exp\left[\log(N) + \log(0.05)/k\right].
\end{aligned}$$

The better approximation is the one with the smaller power in the exponent. Here, these give

$$\begin{aligned}
\hat{n}_1 &\approx \log(0.05)/\log((10000-50)/10000) = 597.65 \\
\hat{n}_2 &\approx 10000 - \exp\left[\log(10000) + \log(0.05)/50\right] = 581.55.
\end{aligned}$$

As $k < n$ using either route, we expect the latter approximation to be the better one. Using simulation, we can solve for the value exactly, getting (according to your book) $n = 581$. This is a useful concept, so I thought I'd hit it one more time.

8.2. The Poisson distribution has been used by traffic engineers as a model for light traffic, based on the rationale that if the rate is approximately constant and the traffic is light (so the individual cars move independently of each other), the distribution of counts of cars in a given time interval or space area should be nearly Poisson (Gerlough and Schuhl 1955). The following table shows the number of right turns during 300 3-min intervals at a specific intersection. Fit a Poisson distribution and test goodness of fit using Pearson's chi-square statistic. Group the last four cells together for the chi-square test. Comment on the fit. It is useful to know that the 300 intervals were distributed over various hours of the day and various days of the week.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13+
Freq.	14	30	36	68	43	43	30	14	10	6	4	1	1	0

Ok, the first step in fitting the Poisson model is simply estimating the Poisson parameter, λ. Using the method of moments, we use

$$\hat{\lambda} = \bar{X} = 1168/300 = 3.893.$$

This is the number of right turns we would expect to see at this intersection in a 3 minute period of light traffic. The next step is to compute expected counts according to the Poisson model to see how well the theoretical model matches reality. To do this, we compute the Poisson probabilities for each cell and multiply by $n = 300$. The matlab function **poisspdf(x,lambda)** is quite useful for this. I've grouped the last 5 cells together so that the expected counts add up to a number greater than 5.

n	0	1	2	3	4	5	6	7	8	9+
O_i	14	30	36	68	43	43	30	14	10	12
E_i	6.11	23.80	46.33	60.13	58.53	45.57	29.57	16.44	8.00	5.51

Note that the last expected count was found by subtracting the sum of the previous ones from 300 so that the total adds up right. Now we compute the contributions of each cell to the goodness of fit total.

n	0	1	2	3	4	5	6	7	8	9+
$\frac{(O_i-E_i)}{\sqrt{(E_i)}}$	3.19	1.27	-1.52	1.02	-2.03	-0.38	0.08	-0.60	0.71	2.78
$\frac{(O_i-E_i)^2}{E_i}$	10.18	1.61	2.30	1.03	4.12	0.15	0.01	0.36	0.50	7.72

Adding these up we get 27.98 as our goodness of fit value. Now we need to determine whether this is too large for our model to be considered a good fit to the data. The distribution of this number should be χ^2_{df}, where the number of degrees of freedom is equal to the number of cells (here 10) less the number of constraints that we force the expected counts to observe. One constraint is simply that the total expected counts must add up to the total observed counts. Beyond that, we are imposing one constraint for every model parameter that we are estimating from the data. In the case of the Poisson model, there is just one parameter, λ, that we are estimating. Hence, the goodness of fit value should follow a $\chi^2_{10-2} = \chi^2_8$ distribution. Checking the table of useful χ^2 thresholds found in your book, we see that $\chi^2_{8,.995} = 21.96$; as our value is greater than this we know that the p-value associated with our result is less than 0.005, so that assuming the model fits would force us to conclude that a very rare event had occurred. As we typically don't want to do this, we would conclude instead that the model did not fit the data well and try to understand

more about why not. In this case, if we check the actual cell contributions, we can see that the fit is worst in the tail-end cells. What this suggests in terms of a refined model is more case-specific and will not be pursued here.

8.6. In an ecological study of the feeding behavior of birds, the number of hops between flights was counted for several birds. For the following data, (a) fit a geometric distribution, (b) find an approximate 95% confidence interval for p, and (c) test goodness of fit.

hops	1	2	3	4	5	6	7	8	9	10	11	12
Freq.	48	31	20	9	6	5	4	2	1	1	2	1

(a)To fit the geometric distribution, we note that if X has a geometric distribution, then $E(X) = 1/p$. Thus, a method of moments estimate of p is simply $\hat{p} = 1/\bar{X}$. Here, $\bar{X} = 363/130 = 2.79$ so $\hat{p} = 0.36$.

(b) Now, to find an approximate 95% confidence interval for p, we need to have some idea as to the variance of our estimate of p. If we can construct standard errors for this, then we can use the CLT to construct approximate confidence intervals. The easiest way to do this is by using maximum likelihood and getting the estimated variance from the curvature of the likelihood function. In this case,

$$\begin{aligned}
L(p) &= \prod_{i=1}^{n} p(1-p)^{x_i-1} \\
l(p) &= \sum_{i=1}^{n} [\log(p) + (x_i - 1)\log(1-p)] \\
&= n\log(p) - n\log(1-p) + \log(1-p)\sum_{i=1}^{n} x_i \\
l'(p) &= \frac{n}{p} + \frac{n}{1-p} - \frac{1}{1-p}\sum_{i=1}^{n} x_i \\
&= \frac{n}{p(1-p)} - \frac{1}{1-p}\sum_{i=1}^{n} x_i
\end{aligned}$$

setting this equal to zero and solving gives $\hat{p} = 1/\bar{X}$, so the method of moments estimator and the mle agree for the geometric model. As for the variance,

$$\begin{aligned}
l''(p) &= -\frac{n}{p^2} + \frac{n}{(1-p)^2} - \frac{1}{(1-p)^2}\sum_{i=1}^{n} x_i \\
&= \frac{n}{p^2(1-p)^2}\left[-(1-p)^2 + p^2 - \bar{x}p^2\right].
\end{aligned}$$

To get the asymptotic standard error of the mle, we evaluate the second derivative at the mle, where $\bar{x} = 1/\hat{p}$, getting

$$\begin{aligned}
l''(\hat{p}) &= \frac{n}{\hat{p}^2(1-\hat{p})^2}\left[-(1-\hat{p})^2 + \hat{p}^2 - \hat{p}\right] \\
&= \frac{n}{\hat{p}^2(1-\hat{p})^2}\left[-1 + 2\hat{p} - \hat{p}^2 + \hat{p}^2 - \hat{p}\right] \\
&= -\frac{n}{\hat{p}^2(1-\hat{p})}.
\end{aligned}$$

The variance of the mle is given by the negative of the inverse of the curvature, so

$$V(\hat{p}) = \frac{\hat{p}^2(1-\hat{p})}{n}$$

and we can form a confidence interval for p as

$$\begin{aligned} 95\%\ CI(p) &= \hat{p} \pm z_{0.975}\sqrt{\frac{\hat{p}^2(1-\hat{p})}{n}} \\ &= 0.36 \pm 1.96\sqrt{\frac{0.36^2(0.64)}{130}} \\ &= (0.31, 0.41). \end{aligned}$$

(c) To test goodness of fit we first compute the expected counts as $n * p * (1-p)^{k-1} = 130 * 0.36 * (0.64)^{k-1}$, getting

hops	1	2	3	4	5	6	7	8	9	10	11	12+
O_i	48	31	20	9	6	5	4	2	1	1	2	1
E_i	46.80	29.95	19.17	12.27	7.85	5.03	3.22	2.06	1.32	0.84	0.54	0.96

At present, the expected counts in the tails are too small for us to trust the χ^2 approximation, so we combine the last 6 cells, getting

hops	1	2	3	4	5	6	7+
O_i	48	31	20	9	6	5	11
E_i	46.80	29.95	19.17	12.27	7.85	5.03	8.93
$\frac{(O_i-E_i)}{\sqrt{(E_i)}}$	0.18	0.19	0.19	-0.93	-0.66	-0.01	0.69
$\frac{(O_i-E_i)^2}{E_i}$	0.03	0.04	0.04	0.87	0.44	0.00	0.48

which, when summed, yields 1.89. For this test, the number of degrees of freedom is 5: 7 (the number of cells), less 1 for the constraint that the expected counts sum to the same total as the observed counts, and less another 1 for the fact that $\hat{p}$ was estimated from the data. Using matlab, **1 - chi2cdf(1.89,5)** gives the p-value of this result as 0.864, so the model appears to mimic reality fairly well.

8.11. In example D of Section 8.4, the method of moments estimate was found to be $\hat{\alpha} = 3\bar{X}$. In this problem you will consider the sampling distribution of $\hat{\alpha}$.

(a) Show that $E(\hat{\alpha}) = \alpha$ – that is, the estimate is unbiased.

This one is fairly straightforward:

$$\begin{aligned} E(\hat{\alpha}) &= E(3\bar{X}) \\ &= 3n^{-1}\sum_{i=1}^{n} E(X_i) \\ &= 3n^{-1}\sum_{i=1}^{n} \alpha/3 = \alpha. \end{aligned}$$

(b) Show that $Var(\hat{\alpha}) = (3 - \alpha^2)/n$.

$$\begin{aligned}
V(\hat{\alpha}) &= V(3\bar{X}) \\
&= 9V(\bar{X}) \\
&= 9n^{-2}\sum_{i=1}^{n} V(X_i). \\
E(X^2) &= \int_{-1}^{1} x^2 \frac{1+\alpha x}{2} dx \\
&= \frac{x^3}{6} + \alpha\frac{x^2}{2}\Big|_{-1}^{1} = \frac{1}{3} \\
V(X) &= E(X^2) - E(X)^2 = \frac{3-\alpha^2}{9} \\
V(\hat{\alpha}) &= 9n^{-1}V(X) = \frac{3-\alpha^2}{n}.
\end{aligned}$$

(c) Use the central limit theorem to deduce a normal approximation to the sampling distribution of $\hat{\alpha}$. According to the approximation, if $n = 25$ and $\alpha = 0$, what is $P(|\hat{\alpha}| > .5)$?

As $\hat{\alpha}$ is the average of a whole bunch of identically distributed terms (it is a scalar multiple of $\bar{X}$), we can use the CLT to claim that the distribution of $\hat{\alpha}$ will be approximately normal, with parameters given by the mean and variance found above:

$$\hat{\alpha} \sim N\left(\alpha, \frac{3-\alpha^2}{n}\right).$$

Thus,

$$\begin{aligned}
P(|\hat{\alpha}| > .5) &= 1 - P(|\hat{\alpha}| < .5) \\
&= 1 - P(-.5 < \hat{\alpha} < .5) \\
&= 1 - P\left(\frac{-.5-\alpha}{\sqrt{(3-\alpha^2)/n}} < \frac{\hat{\alpha}-\alpha}{\sqrt{(3-\alpha^2)/n}} < \frac{.5-\alpha}{\sqrt{(3-\alpha^2)/n}}\right) \\
&= 1 - P(-1.44 < Z < 1.44) = 0.149.
\end{aligned}$$